



ambient
Scientific

AI Needs More Compute. It Cannot Afford More Power.

A Unique Analog-Digital Architecture for Low-Power AI

WHITEPAPER



AI has a power problem

Ask an engineer what limits AI today and you'll rarely hear "math." You'll hear "power." From drones and robotics to hyperscale infrastructure, the same constraint keeps surfacing: power consumption. AI has driven energy demands to unprecedented levels, creating bottlenecks that raise costs and stifle innovation. For the AI platforms of the future, there is no room for compromise. Compute and power efficiency have to go hand in hand, and the organizations that fail to keep up will be left behind.

Ambient Scientific was built to close that gap. Through a unique combination of digital and analog technology, we've created a scalable architecture engineered to meet the demands of modern AI workloads. Our chips are purpose-built for the job, from voice detection to autonomous navigation to foundational AI compute, enabling the critical applications that shape the modern world. Wherever there is AI, there is a need for a compute fabric that scales efficiently from the edge to the cloud. Ambient Scientific harnesses the hybrid power of digital and analog to make that kind of scalable AI possible.

THE PREMISE

Any architecture can add compute for AI. The harder problem is feeding that compute without letting power consumption grow with it.

The real AI bottleneck is memory

AI is often described as a math problem. In practice, it's just as much a memory problem. Most of the compute in a deep-learning graph comes from matrix operations: a neural network layer takes its inputs, multiplies them by trained weights, accumulates the results, and passes the output through additional functions such as biasing, normalization, and activation. These steps repeat across layers until the model produces an inference result. Roughly 95% of that work is dense matrix convolution; the rest is shaping arithmetic such as biasing, normalization, and activation.

Here's where the cost hides. On a conventional single-issue processor, each of those operations requires a sequence of instruction fetches, memory reads, arithmetic operations, and writes. Computing a single neuron with four inputs takes on the order of 38 operations, and roughly three out of every four are memory operations, not math. That overhead compounds fast as models scale. A layer of 32 neurons fed by 32 inputs runs to about 9,000 operations, and stacking

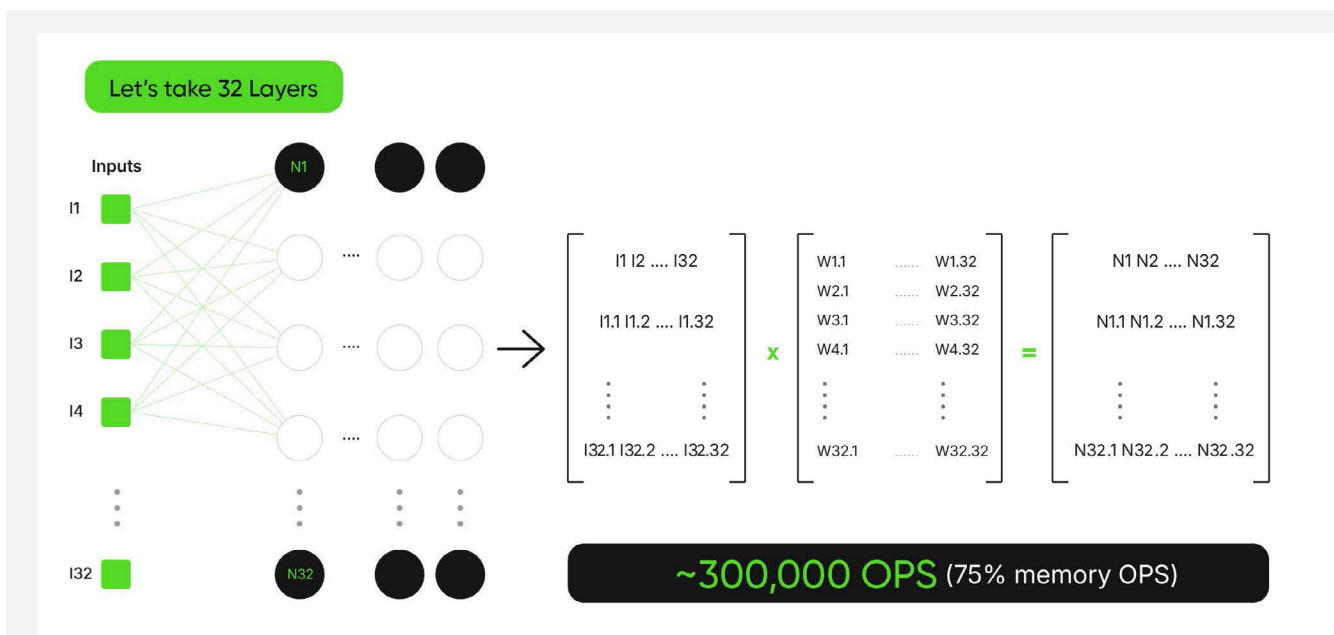


Figure 1. Neural networks are mostly matrix convolutions, and the majority of power consumption and compute goes to memory management. 32 inputs feeding 32 neural nodes: approximately 9,000 operations, 75% of them memory operations.

32 such layers pushes the workload past 300,000 operations, still around 75% memory traffic. Memory is actually where much of the energy goes: moving operands between memory and compute can consume far more energy than the multiply itself, especially when data has to travel off chip.

Adding more multiply-accumulate (MAC) units doesn't solve this on its own. Parallel compute demands parallel access to operands and weights, and conventional multi-ported memory is expensive in area while typically yielding only a couple of MACs per cycle. Add compute without changing the relationship between memory and compute, and the bottleneck simply moves.

Ambient Scientific changed that relationship. To break the bottleneck, we innovated at every level: the circuits, the instruction set, and the software. The result is our digital-analog (A-Cube) architecture, which uses analog where it's efficient and digital where accuracy is critical, reducing energy consumption by orders of magnitude per operation.

A new approach to the matrix

The insight is to let physics do the arithmetic. Using analog pulses, two laws of nature carry the load: Ohm's law multiplies, and Kirchhoff's law sums. The digital side steps in exactly where analog becomes unreliable, handling high-precision accumulation, non-linearities, calibration, and control flow, so nothing is sacrificed for the efficiency gained.

Rather than moving data out of memory and into the compute units, the Ambient Scientific matrix engine brings the compute to the memory. It stores weights in custom SRAM-based structures and performs computation close to where those weights reside, reducing the need to repeatedly fetch them from separate memory and lowering the energy cost of matrix operations. In the current MX8 AI core, the matrix engine delivers up to 256 MAC operations per cycle.

THE ARCHITECTURAL SHIFT

*Don't spend energy shuttling data between memory and compute.
Put the compute where the data already lives in memory.*

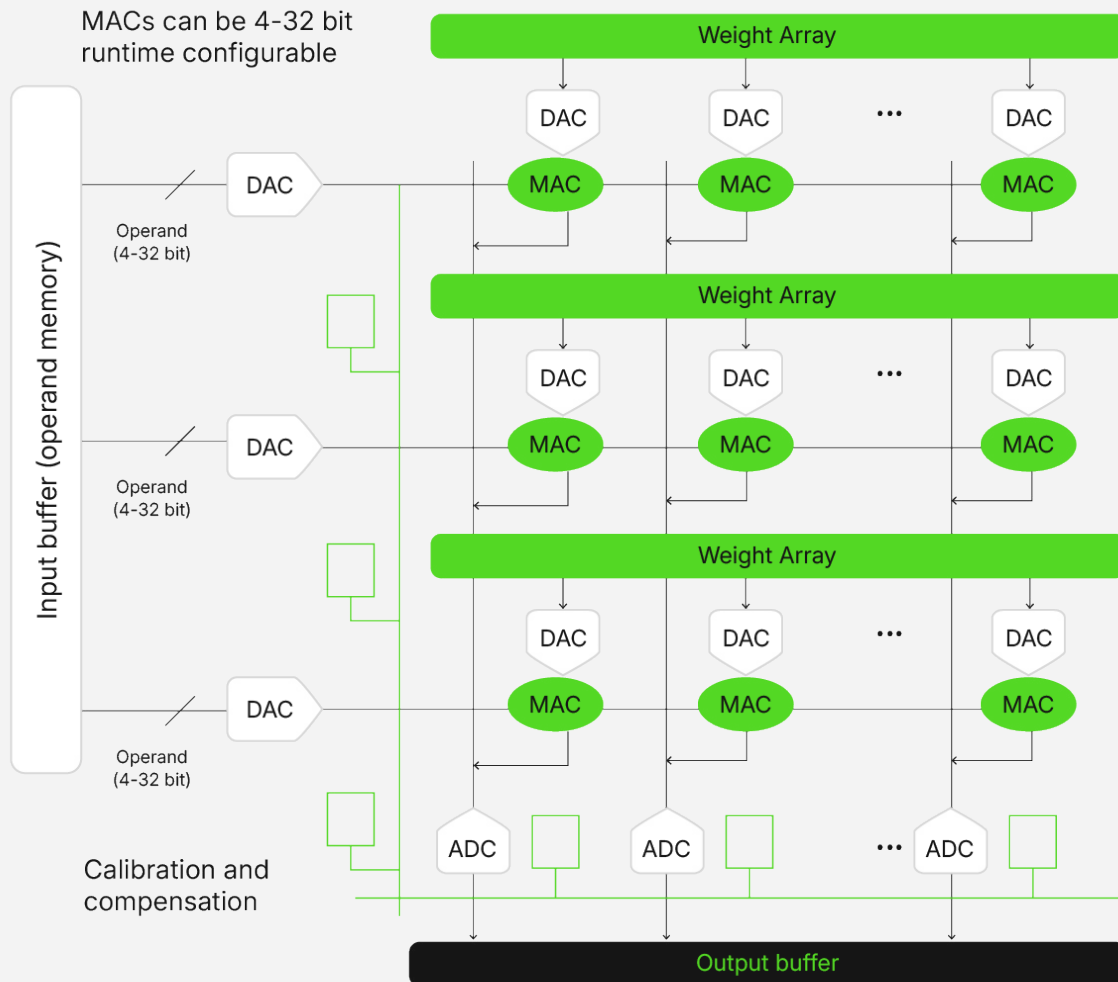
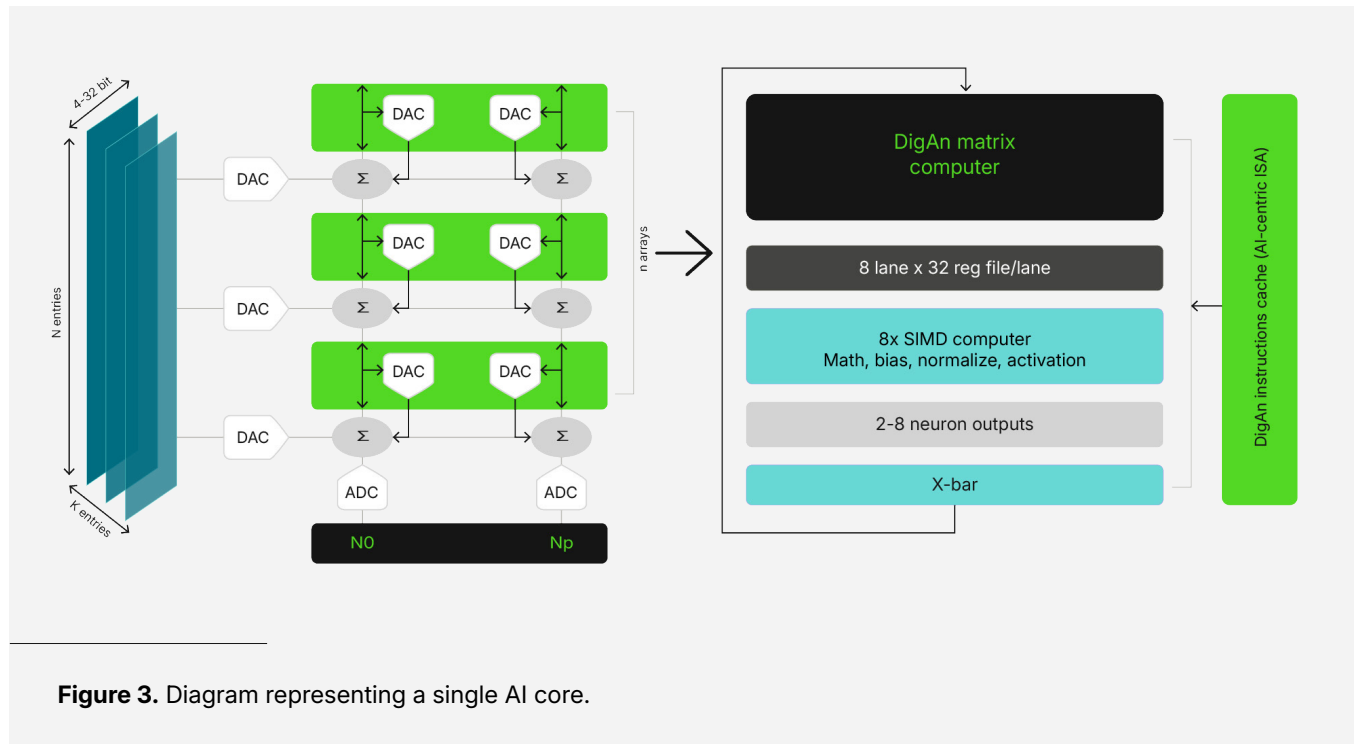


Figure 2. Diagram representing Analog MACs.

A $1 \times 32 \times 8$ matrix operation that would take tens of thousands of cycles on a conventional single-issue core completes in a single cycle on the A-Cube matrix engine. Extend that across 32 layers, and the architecture reduces a workload of more than a million sequential cycles to a small number of matrix-engine cycles.

Scaling from core to engine

Each AI core in the Ambient Scientific architecture includes eight ALUs, each paired with its own dual-port register file. These ALUs handle integer operations at up to 32-bit precision and accelerate common AI operations such as ReLU. From there, data flows to an activation engine that combines digital and analog components: an SRAM array stores a multi-segmented lookup table of coarse activation data points, and mixed-signal logic interpolates between those points to produce a more accurate result.



A crossbar routes data between stages of the core and can feed the output of one neuron or layer back into another, supporting time-series and recurrent topologies. That flexibility lets MX8-based designs support a wide range of model types, including CNNs, RNNs, LSTMs, and GRUs, so a single architecture can target voice, vision, biosignal processing, and sensor fusion without a different accelerator for each use case.

The AI cores run a custom, AI-centric instruction set. Because the matrix engine executes the dense matrix operation in hardware, the instruction stream never has to describe every multiply individually. Instead, it focuses on higher-level control: setting up the next matrix operation, managing post-processing, routing outputs, and moving data through the graph, spending fewer instructions describing arithmetic the hardware already knows how to perform.

The A-Cube architecture is also designed to scale by replication. A single MX8 core can execute neural-network workloads as a self-contained accelerator. Multiple cores connect through a crossbar fabric to form larger compute groups, and those groups compose into larger AI engines. Eight MX8 cores form a micro-cluster, which the software stack treats as an independent deep-learning thread; eight micro-clusters form a cluster; and multiple clusters compose into a full AI engine.

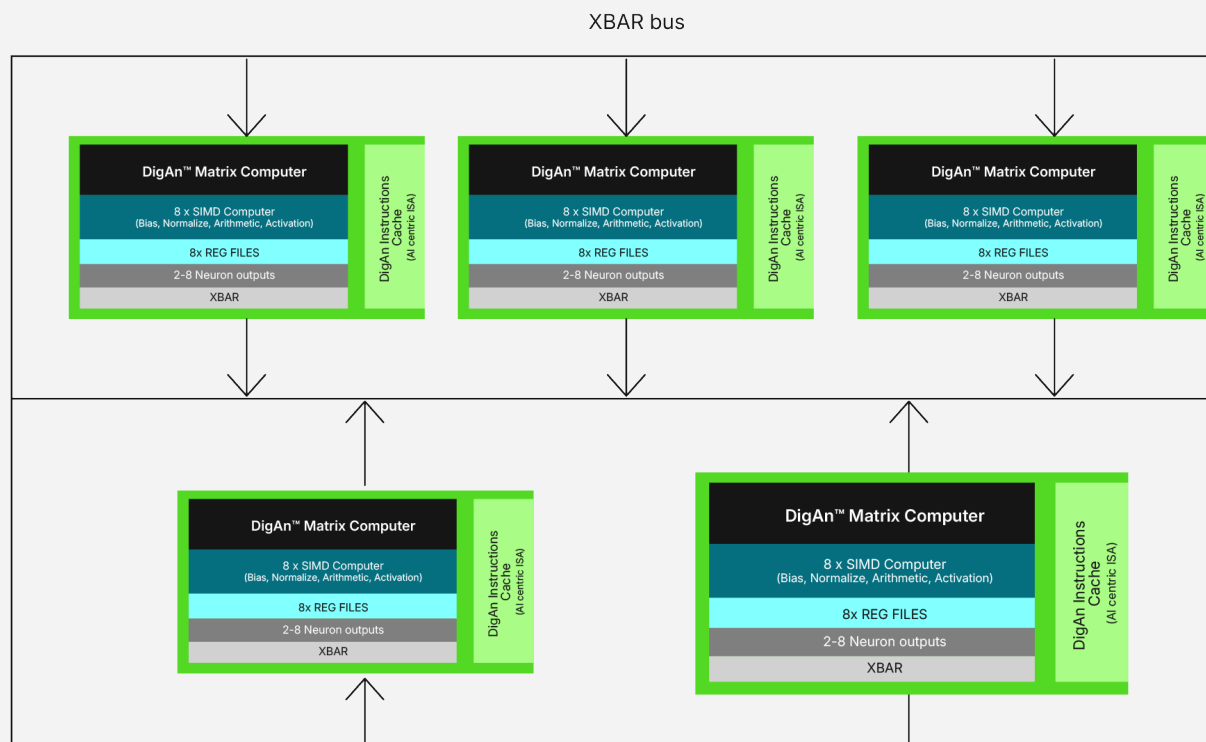


Figure 4. Diagram representing scaling from a single AI core to an AI engine

That structure supports two modes of operation. In one, different cores or clusters run separate AI graphs in parallel, which is ideal for sensor-fusion systems that need to process voice, motion, vision, biosignals, or other inputs at the same time. In the other, the compiler coordinates multiple cores or clusters to execute a single, larger neural-network graph together. Because partitioning is designed into the architecture from the ground up, the compiler can split or fuse graphs efficiently.

Low power at every level

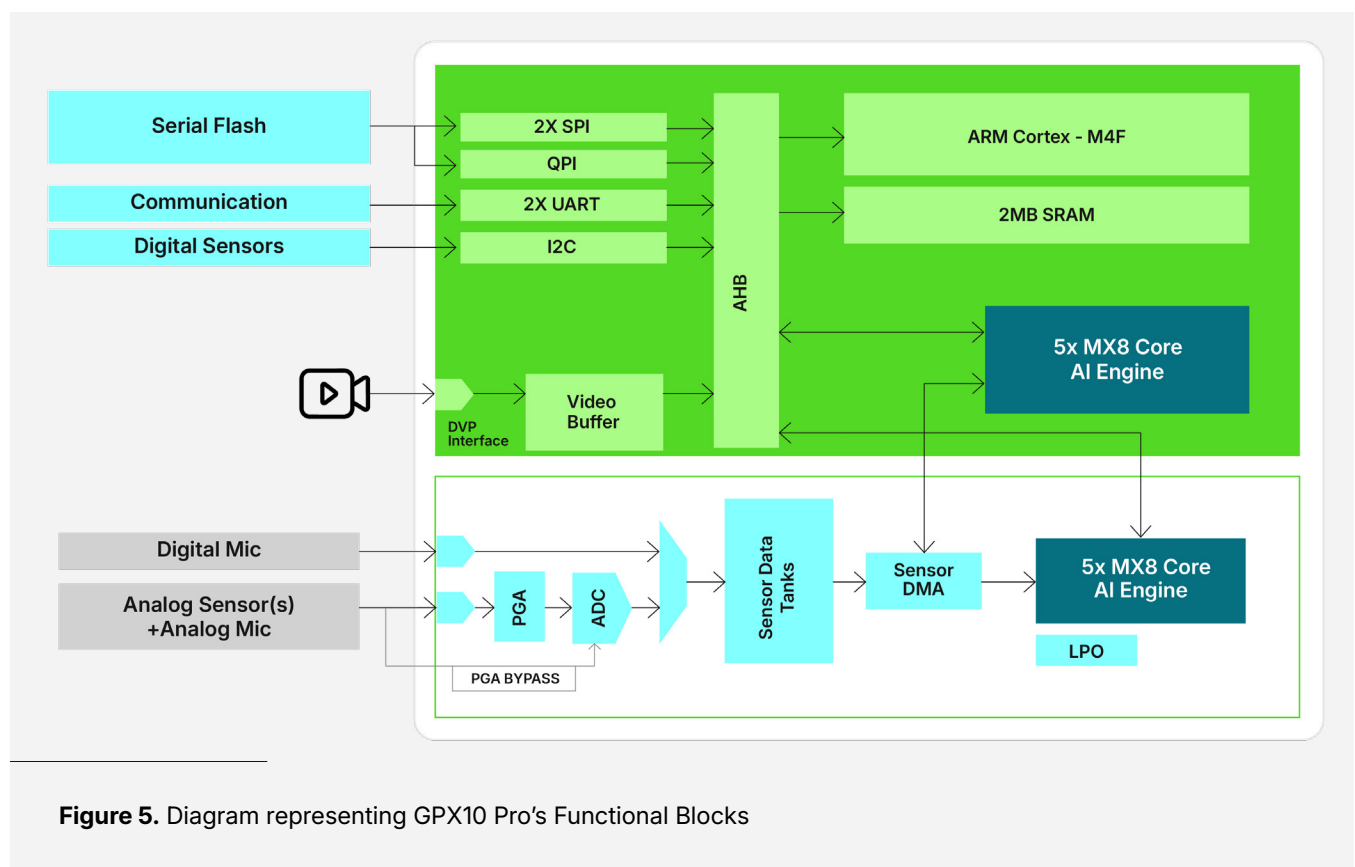
Low-power operation isn't a single feature in the A-Cube architecture. It's built in at every level, from compute and memory design to clocking, data movement, and system partitioning. At the compute level, in-memory matrix operations reduce the need to move operands and weights between distant memory and compute blocks. At the memory level, custom SRAM-based structures and multibank organization keep the matrix engine efficiently fed. At the system level, always-on partitioning lets the most power-hungry parts of the device stay inactive until they're needed.

The memory design earns its efficiency in the details. Memory arrays in the processor subsystem and always-on block use custom SRAM cells, and each array can be paired with custom sense-amplifier circuitry that limits unnecessary bit-line discharge during read operations to lower energy consumption. Multibanked memory arrays continuously feed weights and operands to the MAC units, and the compute engine can load an entire row and multiple columns of a matrix in a single cycle. That cuts memory-read overhead and latency at once, and it's what makes the matrix engine's per-cycle throughput practical to sustain.

The MAC units themselves use analog computation to reduce power relative to conventional digital MAC implementations. Like the SRAMs, they're built from standard CMOS transistors and require no discrete resistors or capacitors, which simplifies the architecture and supports fabrication in standard CMOS processes. The A-Cube matrix block is a full-custom monolithic circuit, designed to hold consistent, calibrated behavior across process, voltage, and temperature (PVT) variation.

Proof in silicon: GPX10 Pro

The concepts behind the A-Cube architecture were validated with Ambient Scientific's first product, the GPX10 Pro, a system-on-chip (SoC) that uses AI-native silicon to deliver high-performance AI inference on battery-powered edge devices. Its AI processing runs across two sets of five MX8 AI cores in two separate power domains, one of which sits in an always-on block that supports ultra-low-power sensor interfacing and fusion. Together, the 10 MX8 cores perform up to 2,560 MAC operations per cycle, for a total peak AI throughput of 512 GOPs.



That compute is backed by 2MB of on-chip SRAM to support large, complex AI models, alongside an Arm® Cortex®-M4F CPU core for classic control functions. Integrated analog functionality includes an ultra-low-power ADC, enhanced I2S logic, and interfaces for up to eight simultaneous analog and multiple digital sensors.

The GPX10 Pro's success brought a scalable roadmap into view, including the upcoming GPX64 and GPX2000. Because Ambient Scientific's AI cores are tile-like and stackable, designs can be customized for optimal performance based on varying workload requirements, so the same foundation extends from always-on edge intelligence to larger AI engines



without a different accelerator architecture at every step. That scalability extends to software as well: the ModelForge software stack connects the hardware to familiar training and deployment flows, letting developers adopt Ambient Scientific technology without rebuilding their AI development process from the ground up or significantly disrupting existing workflows.

Redefining AI compute from Edge to Cloud

The next era of AI won't be defined only by how many operations a processor can perform. It will be defined by how efficiently the system can keep those operations supplied.

Ambient Scientific attacks that problem at every level: bringing computation closer to the memory where data resides, assigning analog and digital circuitry the work each does best, cutting memory traffic, building an AI-centric instruction set, and scaling the result from a single core to a full AI engine.

Together, these choices advance performance, power, and area (PPA), creating a compute architecture built for the realities of AI across the edge-to-cloud continuum, one that treats power efficiency not as a compromise but as a requirement for scale.

Learn more at ambientscientific.ai



ambient
Scientific